NPS55-77-9

NAVAL POSTGRADUATE SCHOOL

Monterey, California

METHODOLOGY FOR EFFICIENCY AND
ALTERATION OF THE LIKELIHOOD SYSTEM

Robert R. Read

March 1977

Approved for public release; distribution unlimited

NAVAL POSTGRADUATE SCHOOL
Monterey, California

Rear Admiral Isham Linder
Superintendent

Jack R. Borsting
Provost

Reproduction of all or part of this report is authorized.

This report was prepared by:

SECURITY CLASSIFICATION OF THIS PAGE (*When Data Entered*)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER NPS-55-77-9	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (*and Subtitle*) METHODOLOGY FOR EFFICIENCY AND ALTERATION OF THE LIKELIHOOD SYSTEM		5. TYPE OF REPORT & PERIOD COVERED Technical Report
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(*s*) Robert R. Read		8. CONTRACT OR GRANT NUMBER(*s*)
9. PERFORMING ORGANIZATION NAME AND ADDRESS Naval Postgraduate School Monterey, California 93940		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Naval Postgraduate School Monterey, California 93940		12. REPORT DATE 15 March 1977
		13. NUMBER OF PAGES 36
14. MONITORING AGENCY NAME & ADDRESS(*if different from Controlling Office*)		15. SECURITY CLASS. (*of this report*) Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (*of this Report*)

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (*of the abstract entered in Block 20, if different from Report*)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (*Continue on reverse side if necessary and identify by block number*)

Efficiency
Estimation
Maximum Likelihood

20. ABSTRACT (*Continue on reverse side if necessary and identify by block number*)

Although maximum likelihood estimates are asymptotically efficient, they are often very hard to find. If this difficulty is caused by some, but not all, of the equations in the system it may be possible to alter the system and make it more manageable. The asymptotic covariance matrix of the new estimate is related to the information matrix. This relationship is characterized and some interpretations are made. Background material on efficiency and lower bounds is included.

METHODOLOGY FOR EFFICIENCY AND ALTERATION OF THE LIKELIHOOD SYSTEM

R. R. Read
Naval Postgraduate School
Monterey, Ca. 93940

ABSTRACT

Although maximum likelihood estimates are asymptotically efficient, they are often very hard to find. If this difficulty is caused by some, but not all, of the equations in the system it may be possible to alter the system and make it more manageable. The asymptotic covariance matrix of the new estimate is related to the information matrix. This relationship is characterized and some interpretations are made. Background material on efficiency and lower bounds is included.

TABLE OF CONTENTS

I. Introduction

This report is one of two reports dealing with some point estimators and their efficiencies for some common probabilistic settings. The companion report [15] is essentially an application's report and contains much numerical work on the asymptotic efficiencies of some estimators that are, in essence, found by modification of the method of moments. The present report provides conceptual background and some computational formulae.

The present report is also a working paper in which old material is presented in a way that reflects my own interests in terms of providing a base for generalization and expansion. Section II contains an overview of the Cramér-Rao type lower bounds for single and multiparameter problems. Hodges' proof [11] of Cramér's basic result [4] is included in order to expose the structure of what happens. Avenues of extension are indicated. In Section III is presented the concept of efficiency and how it extends to multiparameter problems. Although Lemma 3.1 is known [17, p. 378], a proof of it seems hard to locate. The proof by Dan Davis* is presented, again to expose structure and for possible generalization.

The material in Sections IV and V is believed to be new. It deals with the characterization of certain covariance matrices and applies the results to a notion of "directional efficiency." This development is motivated by the fact that although maximum likelihood estimates are asymptotically efficient, they are often extremely hard to find. That is, the system of equations is difficult to solve. Often the fault lies in a few of the equations in the system (rather than all). These few can be replaced in such a way that the resulting system is more easily solved.

---

*Department of Mathematics, USNPS

The resulting estimator will have a covariance matrix related to the information matrix and this relationship is characterized in Theorem 4.1. The loss of efficiency is related to both the number of equations replaced and to the quality of the replacements. The formulae expose the nature of this division. It is shown that there is no loss of efficiency in a subspace of the parameter space.

II. Lower Bounds for the Variance

Consider two sets of random variables $S_1,\ldots,S_k$ and $T_1,\ldots,T_q$, the S's being linearly independent almost surely. That is, there exists no set of constants $a_1,\ldots,a_k$ such that $P\{a_1S_1 + \cdots + a_kS_k = 0\} = 1$ except for all the a's equal to zero. Let A be the covariance matrix of $S_1,\ldots,S_k$ and it follows that A is positive definite. There will be no loss in assuming that each of the S's has expectation zero. Let B be the covariance matrix of the T's and let N be the k by q matrix of cross covariances between the S's and T's.

The following lemma is Hodges' version [11] of Cramér's result [4].

Lemma 2.1. The matrix $B - N'A^{-1}N$ is non-negative definite.

Proof: Let u be any p-vector and v be any q vector. By the Schwarz Inequality

$$\{\mathrm{Cov}(u'S,\ v'T)\}^2 \leq \mathrm{Var}(u'S)\ \mathrm{Var}(v'T) \tag{2.1}$$

where the prime denotes matrix tranpose. Alternatively (2.1) may be written

$$(u'Nv)^2 \leq u'Auv'Bv \tag{2.2}$$

Set $u' = v'N'A^{-1}$ into (2.4) and obtain

$$(v'N'A^{-1}Nv)^2 \leq (v'N'A^{-1}Nv)(v'Bv) \tag{2.3}$$

Since A is positive definite the number $v'N'A^{-1}Nv$ is ≥ 0. If it is positive, one can divide and obtain

$$v'N'A^{-1}Nv \leq v'Bv \tag{2.4}$$

and since B is non-negative definite, the inequality is true also when the left member is zero. Hence for all v we have

$$v'(B - N'A^{-1}N)v \geq 0 \tag{2.5}$$

as required.

The lemma has rather broad value in providing lower bounds. To illustrate the common usage, we introduce the setting for a case of regular point estimation. Let the parameter space Θ be an open subset of p-dimensional Euclidean space and let the population sampled have a (generalized) density function $f(x;\theta)$, $\theta \in \Theta$. Given a sample $x_1,\ldots,x_n$ of size n the likelihood function will be denoted by

$$L(x;\Theta) \equiv f(x_1;\theta)\ f(x_2;\theta),\ldots,\ f(x_n;\theta)\ . \tag{2.6}$$

Moreover, the quantities

$$s_r(x) = \partial \ln f(x;\theta)/\partial\theta_r\ , \qquad r = 1,\ldots,p \tag{2.7}$$

$$s_{r,j}(x) = \partial^2 \ln f(x,\theta)/\partial\theta_r\ \partial\theta_j, \quad r,j = 1,\ldots,p \tag{2.8}$$

are assumed to exist.

The setting for a case of regular point estimation requires assumptions concerning the ability to differentiate under the integral sign. Following Wilks [17] we assume (using $F(x) = \int_{-\infty}^{x} f(u;\theta)\, du$)

$$E\{s_r(X)\} = \frac{\partial}{\partial\theta_r} \int dF(x;\theta) = 0; \quad r = 1,\ldots,p \tag{2.9}$$

$$E\{s_r(X)\, s_j(X)\} + E\{s_{r,j}(X)\} = \frac{\partial^2}{\partial\theta_r\, \partial\theta_j} \int dF(x;\theta) = 0 \tag{2.10}$$

$$r,j = 1,\ldots,p$$

From (2.9) and (2.10) it follows that the covariance matrix of the $\{s_r(X)\}$ can be calculated from

$$\mathrm{Cov}\{s_r(X),\, s_j(X)\} = -E\{s_{r,j}(X)\} \tag{2.11}$$

The symbol Λ will be reserved for this matrix. It is the (Fisher) Information matrix.

A common choice for $S_1,\ldots,S_k$ of Lemma (2.1) is

$$S_r = \sum_{i=1}^{n} s_r(X_i)\ , \qquad r = 1,\ldots,k \leq p \tag{2.12}$$

Then the S's are also functions of θ. The linear independence assumption and hence the positive definite nature of A must hold for almost all θ, and

$$A = n\Lambda \tag{2.13}$$

The random variables $T_1,\ldots,T_q$ are statistics and suppose they are used to estimate $\theta_1,\ldots,\theta_q$. If we choose not to require that the T's be unbiased estimators, then we should characterize the risk matrix

$$R = E\{(T-\theta)(T-\theta)'\} = B + b \tag{2.14}$$

where, in (2.14), $\theta' = (\theta_1,\ldots,\theta_q)$ and b is the matrix of products of bias terms

$$b = (\mu-\theta)(\mu-\theta)' \tag{2.15}$$

using $\mu = E(T)$. Since b is obviously non-negative definite it follows that Lemma 2.1 could have stated that $R - N'A^{-1}N$ is non-negative definite, but this is not as sharp.

First Application. Let $p = q = k = 1$ and assume that T is a regular estimate for θ in the sense that $E(T)$ can be differentiated under the integral sign. Then

$$\begin{aligned} N = E(TS) &= \int T(x) \frac{\partial \ln L(x;\theta)}{\partial\theta} L(x;\theta)\, dx \\ &= \frac{\partial}{\partial\theta} \int T(x)\, L(x;\theta)\, dx = \frac{\partial}{\partial\theta} E(T) = 1 + b'(\theta) \end{aligned} \tag{2.16}$$

Then, applying Lemma 2.1 and using (2.13) and (2.16) provides the familiar lower bound for the variance of T

$$\sigma_T^2 \geq \frac{[1 + b'(\theta)]^2}{n\Lambda} \tag{2.17}$$

There are common examples such that σ_T^2 equals the lower bound. Then T can be said to be the best regular estimator.

Second Application. Suppose $q = 1$, $k \leq p$, and T is intended to estimate θ_1. Also T is regular in the sense that it can be differentiated under the integral sign with respect to each $\theta_1,\ldots,\theta_k$. Thus, the rth element of N is

$$N_r = E\{TS_r\} = \int T(x) \frac{\partial L(x;\theta)}{\partial\theta_r} dx$$

$$= \frac{\partial}{\partial\theta_r} \int T(x)\, L(x;\theta)\, dx = \delta_{1r} + b_r(\theta) \ , \quad r = 1,\ldots,k \ . \qquad (2.18)$$

where δ_{1r} is the Kronecker delta and b_r is the partial derivative of the bias function with respect to θ_r. The resulting inequality is

$$\sigma_T^2 \geq N'A^{-1}N \qquad (2.19)$$

where A is the upper left k by k corner of $n\Lambda$. This provides a lower bound for the variance (risk) of an estimator for θ_1 when there are nuisance parameters present. Moreover the bound is nondecreasing as k is increased. This has importance because it affects the sharpness of the bound. Let us justify this point by drawing attention to relationships with the multiple correlation coefficient.

Consider the projection (in L_2) of T on the subspace spanned by $1, S_1,\ldots,S_k$. The mathematical problem is to choose the scalar c and the k vector β so as to minimize $E\{T - c - \beta'S\}^2$. The solution is $c = E(T) = \theta_1 + b(\theta)$ and $\beta' = N'A^{-1}$. The projection is

$$\hat{T} = E(T) + N'A^{-1}S \qquad (2.20)$$

and easy calculations show the mean square error

$$MSE = E\{T-\hat{T}\}^2 = \mathrm{Var}(T) - N'A^{-1}N \tag{2.21}$$

which is surely nonincreasing as k increases. Working next on an inner product term produces

$$\begin{aligned} E\{(T-\hat{T})(\hat{T}-\theta_1)\} &= E\{(T-\hat{T})(b + \beta'S)\} \\ &= \beta'E(ST) - \beta'E(S\hat{T}) = \beta'N - \beta'ES(\mu + \beta'S) \\ &= N'A^{-1}N - N'A^{-1}AA^{-1}N = 0 \end{aligned} \tag{2.22}$$

It follows that the orthogonal decomposition

$$E(\hat{T}-\theta_1)^2 = E\{T-\hat{T}\}^2 + E(\hat{T}-\theta_1)^2 \tag{2.23}$$

is valid. Similar calculations yield

$$\begin{aligned} E\{\hat{T}-\theta_1\}^2 &= \mathrm{Var}(\hat{T}) + b^2(\theta) \\ &= N'A^{-1}N + b^2(\theta) \end{aligned} \tag{2.24}$$

Using (2.21) and (2.24) allows (2.23) to be rewritten

$$R_T = MSE + b^2(\theta) + N'A^{-1}N \tag{2.25}$$

Since R_T and $b(\theta)$ are fixed, the nonincreasing feature of MSE implies that the bound $N'A^{-1}N$ is nondecreasing.

The quantity ρ defined by

$$1 - \rho^2 = \frac{MSE + b^2(\theta)}{\mathrm{Var}(T) + b^2(\theta)} \tag{2.25}$$

plays a role analogous to the multiple correlation coefficients. Clearly $0 \leq \rho^2 \leq 1$.

Third Application. Suppose $p = q = 1$ and, following Bhattacharyya [2], we define

$$S_r = \frac{1}{L(x;\theta)} \frac{\partial^r L(x;\theta)}{\partial \theta^r} , \qquad r = 1,\ldots,k \tag{2.26}$$

Now the regularity conditions must be extended to include k differentiatio under the integral sign so that

$$E(S_r) = \frac{\partial^r}{\partial \theta^r} \int L(x;\theta)\, dx = 0 . \tag{2.27}$$

Also $E(T)$ must be differentiated under the integral k times so we can use

$$\begin{aligned} N_r = E(TS_r) &= \frac{\partial^r}{\partial \theta^r} \int T(x)\, L(x;\theta)\, dx \\ &= \frac{\partial^r}{\partial \theta^r} (\theta + b(\theta)) \\ &= \delta_{1r} + b^{(r)}(\theta) \end{aligned} \tag{2.28}$$

Further

$$A_{rj} = E \left\{ \frac{1}{L^2(x;\theta)} \frac{\partial^r L(s;\theta)}{\partial \theta^r} \frac{\partial^j L(x;\theta)}{\partial \theta^j} \right\} \tag{2.29}$$

This technique can produce an increasing sequence of bounds for some problem and a sequence of constants for others [2,11].

The preceding description is more general than one usually finds. Typically the T's are the estimators (hence statistics). The S's are functions of the model. The bound is trivial if the T's are not correlated with the S's, i.e. $N = 0$. Thus there is a hierarchy of problems which, in vague terms, might be expressed as follows: Given S find the "best" T among all T's correlated with S. Having characterized this problem for each choice of S, find the "best" S.

Some progress has been made in this approach to finding good estimators. It usually appears under the name of minimum variance unbiased regular estimation. The bounds are sharp in some of the more popular settings. An example is included in a setting for which sharp bounds are not known.

Example (Geometric). Consider a sample of size one from a geometric density

$$f(x;p) = pq^x \qquad \text{for} \quad x = 0,1,2,... \tag{2.30}$$

This is a member of the Darmois-Koopman family. It is known [11, p. 2-21] that

$$\begin{aligned} T &= 1 \qquad \text{if} \quad X = 0 \\ &= 0 \qquad \text{if} \quad X > 0 \end{aligned} \tag{2.31}$$

is the only unbiased estimator for p. Its variance pq certainly is uniformly minimum among all unbiased estimators. Applying (2.17) and (A.9) from the appendix produces the classical Cramér-Rao lower bound for regular unbiased estimators

$$L = qp^2 \tag{2.32}$$

Clearly this is not sharp ($< pq$). The margin is made graphic in Figure 2.1.

The maximum likelihood estimator, $\hat{p} = (1+x)^{-1}$ is biased. Its variance and bias function are developed in the appendix, see (A.4) to (A.6). Its risk is

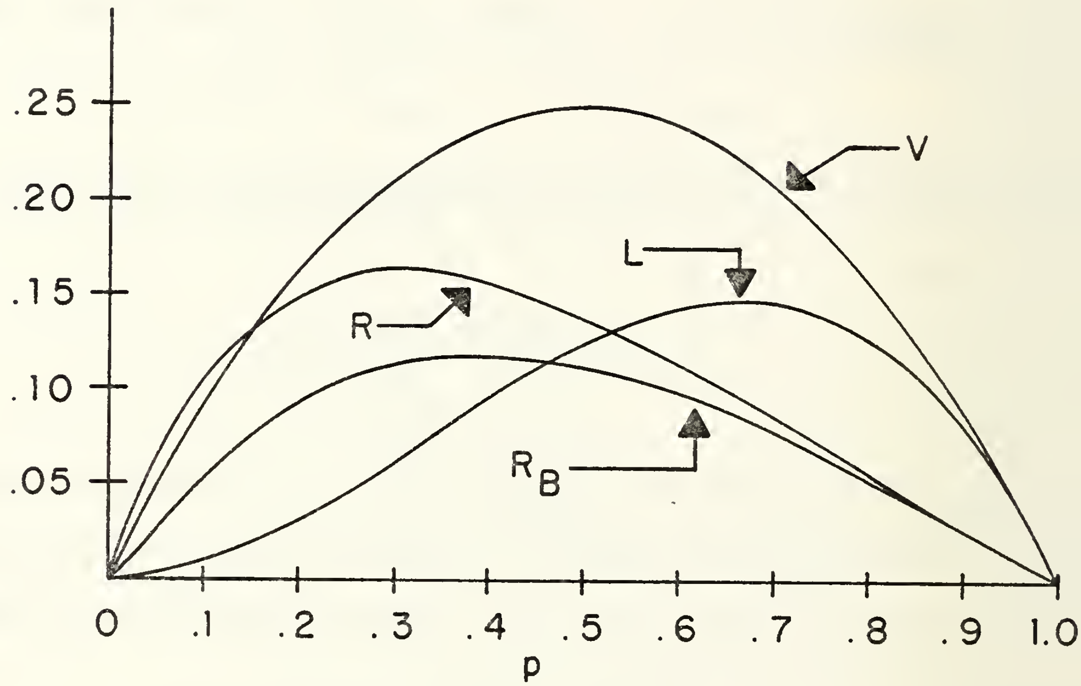
$$R = \sigma^2(\hat{p}) + b^2(p) \tag{2.33}$$

and, using (2.17) and (2.32), a lower bound for the risk is

$$R_B = qp^2(1 + b'(p))^2 + b^2(p) \tag{2.34}$$

and $b'(p)$ is given in (A.7). Both R and R_B are included in Figure 2.1.

Figure 2.1
Comparison of Risks and Bounds Geometric Distribution



L = Cramér-Rao Lower Bound

V = Variance of the Unbiased Estimator

R = Risk of the Maximum Likelihood Estimator

R_B = Lower Bound for the Risk.

Since $R > R_B$ a sharp bound still has not been found. There remains the possibility that another estimator, with a different bias function, may have a variance that matches its lower bound (2.17). It is curious that $\hat{p}$ is not uniformly better than T.

Returning to unbiased estimators, the classical Cramér-Rao bound can be improved by the method of Bhattachrayya. Here Λ is the covariance matrix of the $\{S_r\}_1^k$, of (2.26) and N is a vector of zeros except for unity in the first position, by (2.28). By (2.19) the lower bound is

$$L_k = N'\Lambda^{-1}N \tag{2.35}$$

and these are nondecreasing in k. In [11, p. 2-23] it is shown that

$$\begin{aligned} L_1 &= p^2q \quad \text{(Cramér-Rao)} \\ L_2 &= p^2q(1 + q) \end{aligned} \tag{2.36}$$

Attention has been brought to the notion that if $L_k = p^2q(1 + q + \cdots + q^{k-1})$ then the sequence $\{L_k\}$ would converge to pq and the bound would be sharp. It is shown in the appendix that L_3 does possess the form speculated.

III. Concept of Efficiency

When two estimates of the same quantity have unequal risks, the ratio of the smaller risk to the larger one is called the efficiency of the latter estimator with respect to the former one. The usefulness of this measure presumes that the distributions of the two statistics have roughly similar shapes and the risks are (approximately) inversely proportional to the sample size. Thus the efficiency can be interpreted as the ratio of sample sizes needed so that the two statistics could estimate the parameter equally well.

Often the Cramér-Rao lower bound is used in place of an estimat to serve as the standard of comparison. This is satisfactory provided the bound is sharp.

A useful extension of the idea of efficiency to the multiple parameter case appears in [5] and utilizes the concept of an ellipse of concentration. To explain, let $p = 2$ and consider an estimator $(T_1,$ of (θ_1,θ_2) and let B be the covariance matrix as before. The ellip of concentration is that ellipse in the plane centered at (θ_1,θ_2), whi serves as the positive sample space of uniformly distributed random var ables (U_1,U_2) that have the same covariance matrix B. The efficacy this lies in the standardization of geometrical shape. The comparison two estimators (T_1,T_2) and (T_1',T_2') is accomplished by comparing the ellipses of concentration.

Convariance matrices are often (roughly) inversely proportional sample size. Then the concept of efficiency being the ratio of sample required to perform the same job is preserved if the determinants of th covariance matrices are compared, that is, the squared area of the elli of concentration. This extends obviously to the general multiparameter case, the ratio of determinants being the ratio of squared contents of hyper-ellipsoids of concentration.

The following lemma is useful in defining multiparameter effici with respect to a standard.

<u>Lemma 3.1</u>. Let $q = p$ and B (as well as A) be positive definite. If $u'Auv'Bv \geq (u'v)^2$ for all u, v, then $|AB| \geq 1$ (where the vertica denote determinant).

<u>Proof</u>: (Dan Davis). Let P be a similarity transformation that diagonalizes A, $(P'AP = D,\ P'P = I)$. The hypothesis becomes (upon replacing u with Pu)

$$u'Duv'Bv \geq (u'P'v)^2$$

and in this expression, let us also replace v with Pv, yielding

$$u'Duv'P'BPv \geq (u'v)^2 \tag{3.1}$$

Since A is positive definite, the diagonal elements of D are positive and $D^{-1/2}$ exists. Let $u = D^{-1/2}w$ and $v = D^{1/2}z$. Thus (3.1) becomes

$$w'wz'D^{1/2}P'BPD^{1/2}z \geq (w'z)^2 \tag{3.2}$$

Let $C = D^{1/2}P'BPD^{1/2}$ and note that

$$|C| = |D^{1/2}||D^{1/2}||B| = |A||B| = |AB| \tag{3.3}$$

and the proof will be completed when we have shown that all the eigenvalues of C are ≥ 1. The inequality (3.2) is preserved if C is rotated to diagonal form. The diagonal elements are the eigenvalues. Let $w = z = \{\delta_{ij}\}_{j=1}^{p}$, that is, the unit vector in the ith component direction. It follows from (3.2) that the ith eigenvalue of C is ≥ 1. This is true for each $i = 1,\ldots,p$. Q.E.D.

The hypothesis of Lemma 3.1 is not always met in a setting of regular estimation. Let us examine this question for regular estimators. Consider the vector equation

$$E\{T\} = \theta + b(\theta) \tag{3.4}$$

where b is the vector of bias functions. Taking the partial derivative of the ith component of (3.4) with respect to θ_r produces

$$\frac{\partial}{\partial\theta_r} \int T_i(x)\ L(x;\theta)dx = \int T_i(x)\ S_r L(x;\theta)dx = \delta_{ir} + \frac{\partial b_i}{\partial\theta_r} \tag{3.5}$$

or

$$N = \mathrm{Cov}(S,T) = I + \left\{ \frac{\partial b_i}{\partial\theta_r} \right\} \tag{3.6}$$

Since (2.2) is valid in general, the hypothesis of Lemma 3.1 is met when T is unbiased, and also in some cases when the bias functions decrease in appropriate ways.

For cases of unbiased regular estimation, we can set $A = n\Lambda$ and define the efficiency of the multiparameter estimator T to be

$$\mathrm{Eff}(T) = \frac{1}{n|B\Lambda|}\ . \tag{3.7}$$

IV. Asymptotic Covariance Matrices

We begin with two lemmas that are wholly mathematical in nature.

Suppose the two quadratic forms A and B are both positive defin and have certain submatrices in common. Use row and column permutations, if necessary, and assume the partitioned form,

$$A = \begin{bmatrix} E & F \\ F' & G \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} E & F \\ F' & H \end{bmatrix} \tag{4.1}$$

where E is q by q, G is p-q by p-q, F is q by p-q. This structure is most useful when (4.1) is the most extreme such representation, that is, no row of F', G is equal to any row of F', H.

Lemma 4.1. The rank of $A^{-1} - B^{-1}$ is $\leq$ p-q.

Proof: The expressions for inverting partitioned matrices [7, p. 165] allow the representations

$$A^{-1} = \begin{bmatrix} X & -E^{-1}FW \\ -G^{-1}F'X & W \end{bmatrix}, \quad B^{-1} = \begin{bmatrix} Y & -E^{-1}FZ \\ -H^{-1}F'Y & Z \end{bmatrix} \tag{4.1'}$$

where

$$\begin{aligned} X &= [E-FG^{-1}F']^{-1} = E^{-1} + E^{-1}FWF'E^{-1} \\ Y &= [E-FH^{-1}F']^{-1} = E^{-1} + E^{-1}FZF'E^{-1} \\ W &= [G-F'E^{-1}F]^{-1} = G^{-1} + G^{-1}F'XFG^{-1} \\ Z &= [H-F'E^{-1}F]^{-1} = H^{-1} + H^{-1}F'YFH^{-1} \end{aligned} \tag{4.2}$$

and E, G, H, W, X, Y, and Z are symmetric.

Using the form

$$A^{-1} - B^{-1} = \begin{bmatrix} X-Y & -E^{-1}F(W-Z) \\ -G^{-1}F'X+H^{-1}F'Y & W-Z \end{bmatrix} \tag{4.3}$$

note that the first q rows of (4.3) will be represented as a linear combination of the last p-q rows, as soon as it is exhibited that

$$E^{-1}F[G^{-1}F'X-H^{-1}F'Y] = X-Y \ . \tag{4.4}$$

Using the symmetry of A^{-1} and B^{-1}, (4.1), the left member of (4.4) can be written as

$$E^{-1}F(W-Z)F'E^{-1} \tag{4.5}$$

and the fact that this is X-Y can be seen by subtracting the right members of the first two expressions in (4.2). Thus the rank of $A^{-1} - B^{-1}$ is no more than p-r.

Lemma 4.2. The rank of $A^{-1} - B^{-1}$ is p-r if G-H is invertible.

Proof: It is sufficient to show that $(W-Z)^{-1}$ exists. Using the ide

$$(W-Z) = Z(Z^{-1}-W^{-1})W$$

it is seen that $(W-Z)^{-1} = W^{-1}(Z^{-1}-W^{-1})^{-1}Z^{-1}$, or, using (4.2)

$$[G-F'E^{-1}F][H-G]^{-1}[H-F'E^{-1}F]$$

which exists if H-G is invertible.

The structure treated above occurs when $A = \Lambda$, the informati matrix, and $B = M$ where $\frac{1}{n} M + o(\frac{1}{n})$ is the covariance matrix of est resulting from a system that has some but not all of the likelihood equations. We are dealing with the asymptotic forms.

It is assumed that the estimation equations have the form

$$g(x,\theta) = 0$$

where $g = (g_1,\ldots,g_p)$ and each component is a symmetric function of i.e. is invariant under permutations of $x_1,\ldots,x_n$. In order for the (4.8) to have a unique solution $\tilde{\theta}$, it is necessary (by the Implicit F Theorem [16]) that the Jacobian

$$J = \left|\frac{\partial g_1 \cdots g_p}{\partial \theta_1 \cdots \theta_p}\right| \neq 0$$

Moreover, we require

$$E\{g(x,\theta)\} = 0$$

a property acquired by manipulation.

For convenience of analysis, it is further assumed that the es equations (4.8) have been scaled so that

$$\mathrm{Var}\{g_j(x,\theta)\} = \frac{c_j}{n} + o(\frac{1}{n}) \tag{4.10}$$

for some positive constants $c_1,\ldots,c_p$. Moreover, for our purposes is it assumed that the $\{g_j(x,\theta)\}$ have bounded continuous partial derivatives with respect to $\theta_1,\ldots,\theta_p$ and that $\tilde{\theta}$, resulting from the solution of (4.8) is consistent. Hence the estimate is asymptotically unbiased.

Reference [10] contains a deep treatment of the general question of the existence of consistent, asymptotically normal estimates. There, the functions $\{g_j\}_1^p$ are averages of the form $\frac{1}{n}\sum_{i=1}^n g_j(x_i,\theta)$ and this structure precludes much of what has been assumed so far. Indeed, all of the examples treated thus far can be cast in this average form. The goals of the present work are much less pure than those of LeCam, and the question of verifying the consistency of $\tilde{\theta}$ is left to the applier.

It is noted that the equations for maximum likelihood estimation can be cast into this structure.

Finally, let $A(x,\theta)$ be the p by p matrix of partial derivatives $\{\partial g_i/\partial\theta_k\}$ and assume that

$$A_{ik}(x,\tilde{\theta}) \to E\left(\frac{\partial g_i}{\partial\theta_k}\right) \qquad \text{a.s.} \tag{4.11}$$

as $n \to \infty$, and the resulting limit matrix will be denoted by $A = A(\theta)$.

The assumptions allow the first order expansion

$$g(x,\theta) = g(x,\tilde{\theta}) + A(x,\ \tilde{\theta} + \rho(\theta-\tilde{\theta}))\ (\theta-\tilde{\theta}) \tag{4.12}$$

where ρ is a diagonal matrix of random numbers belonging to the interval [0,1]. Since the system is soluble, $g(x,\tilde{\theta}) = 0$ and we can write

$$(\tilde{\theta}-\theta) = -A^{-1}(x, \tilde{\theta} + \rho(\theta-\tilde{\theta}))\ g(x,\theta) \qquad (4.$$

The continuity of A implies that of g and of A^{-1}. Letting the asymp covariance matrices be defined by

$$M = \text{limit } nE(\tilde{\theta}-\theta)(\tilde{\theta}-\theta)' \qquad (4.$$

$$C = \text{limit } nE\{g(x,\theta)\ g'(x,\theta)\} \qquad (4.$$

it follows that

$$M = A^{-1}C(A^{-1})' \qquad (4.$$

When $g = 0$ is the set of likelihood equations, i.e. from (2.12)

$$\frac{1}{n} S_r = 0 \quad \text{for} \quad r = 1,\ldots,p \qquad (4.$$

it is well-known (and easily verified) that M of (4.14) and (4.16) is (the inverse of the information matrix), C of (4.15) is Λ, and A of (4.11) is, by (2.11),

$$E\{s_{ik}(X)\} = -\Lambda \qquad (4.$$

which is symmetric in this case.

Now let us suppose, without loss of generality, that only the fir q equations of (4.17) are used in the system $g = 0$. Let us denote this subset by the symbols $\mu = 0$, and let the remaining p-q equations be $h =$ Thus, in partitioned form, (4.8) becomes

$$g = \{{\mu \atop h}\} = 0 \qquad (4.$$

All assumptions are met and we can proceed formally

$$C = \text{limit } nE\{gg'\} = \text{limit } n \begin{pmatrix} E(\mu\mu') & E(\mu h') \\ E(h\mu') & E(hh') \end{pmatrix} = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix} \qquad (4.$$

The information matrix can be partitioned similarly

$$\Lambda = \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix} \tag{4.21}$$

where Λ_{11} is a q by q matrix. The following result is obvious:

<u>Lemma 4.3.</u> $C_{11} = \Lambda_{11}$.

Let us define a p-q by q matrix g_{21} whose elements are $E\{\partial h_j/\partial\theta_k\}$ for $j = 1,\ldots,p-q$ and $k = 1,\ldots,q$.

<u>Lemma 4.4.</u> $C_{21} = -g_{21}$.

<u>Proof:</u> From (4.20) it is seen that the (j,k)th element of C_{21} is the limit of

$$E\sum_{i=1}^{n}\left\{\frac{\partial \ln f(X_i,\theta)}{\partial\theta_k} h_j(X,\theta)\right\} = nE\left\{\frac{\partial \ln f(X,\theta)}{\partial\theta_j} h_j(X,\theta)\right\}$$

by the interchangeability property of the function h_j with respect to the $X = (X_1,\ldots,X_n)$. Now the assumptions imply that the equation $0 = E[h_j(X,\theta)]$ $= \int h_j(x,\theta)\exp\{\sum_{i=1}^{n} \ln f(x_i,\theta)\}$ can be differentiated with respect to each θ; under the integral sign. So doing produces

$$0 = E\left\{\frac{\partial h_j(X,\theta)}{\partial\theta_k}\right\} + E\left\{\sum_{i=1}^{n} \frac{\partial \ln f(x_i,\theta)}{\partial\theta_k} h_j(X,\theta)\right\}$$

which is the desired result.

Lemmas 4.3 and 4.4 provide the representation

$$C = \begin{pmatrix} \Lambda_{11} & -g_{21}' \\ -g_{21} & C_{22} \end{pmatrix} \tag{4.22}$$

for the asymptotic covariance matrix of $g' = (\mu', h')$. It follows rather quickly from (4.11) and the discussion accompanying (4.18) that

$$A = \begin{pmatrix} -\Lambda_{11} & -\Lambda_{12} \\ g_{21} & g_{22} \end{pmatrix} \tag{4.23}$$

where g_{22} is defined to complement g_{21} and has elements $E\{\partial h_j/\partial\theta_k\}$ for $j = 1,\ldots,p-q$ and $k = q+1,\ldots,p$.

Using (4.22) and [7] we can characterize C^{-1} as

$$C^{-1} = \begin{pmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{pmatrix} \tag{4.24}$$

where, using $g_{12} = g_{21}'$,

$$G_{11} = [\Lambda_{11} - g_{12}C_{22}^{-1}g_{21}]^{-1} = [I + \Lambda_{11}^{-1}g_{12}G_{22}g_{21}]\Lambda_{11}^{-1} \tag{4.25}$$

$$G_{12} = G_{21}' = \Lambda_{11}^{-1}g_{12}G_{22} = G_{11}g_{12}C_{22}^{-1} \tag{4.26}$$

$$G_{21} = G_{12}' = C_{22}^{-1}g_{21}G_{11} = G_{22}g_{21}\Lambda_{11}^{-1} \tag{4.27}$$

$$G_{22} = [C_{22} - g_{21}\Lambda_{11}^{-1}g_{12}]^{-1} = C_{22}^{-1}[I + g_{21}G_{11}g_{12}C_{22}^{-1}] \tag{4.28}$$

Theorem 4.1. If the first q equations of $g = 0$ are likelihood equations, then

$$M^{-1} = \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & H \end{pmatrix}$$

where

$$H = \Lambda_{21}G_{11}\Lambda_{12} - \Lambda_{21}G_{12}g_{22} - (\Lambda_{21}G_{12}g_{22})' + g_{22}G_{22}g_{22}$$

Proof: From (4.16) we have

$$M^{-1} = A'C^{-1}A \tag{4.29}$$

Multiplying the partitioned matrices (4.23) and (4.24) followed by applying the relationships (4.25) and (4.26) yields the product form

$$C^{-1}A = \begin{pmatrix} -I & G_{12}g_{22} - G_{11}\Lambda_{12} \\ 0 & G_{22}g_{22} - G_{21}\Lambda_{12} \end{pmatrix} \tag{4.30}$$

and draws attention to the relationships

$$G_{11}\Lambda_{11} - G_{12}g_{21} = I \tag{4.31}$$

$$G_{21}\Lambda_{11} - G_{22}g_{21} = 0 \tag{4.32}$$

Multiplying (4.30) on the left by A' from (4.23) produces the intermediate form

$$\begin{pmatrix} \Lambda_{11} & \Lambda_{11}[G_{11}\Lambda_{12} - G_{12}g_{22}] + g_{12}[G_{22}g_{22} - G_{21}\Lambda_{12}] \\ \Lambda_{21} & \Lambda_{21}[G_{11}\Lambda_{12} - G_{12}g_{22}] + g_{22}[G_{22}g_{22} - G_{21}\Lambda_{12}] \end{pmatrix}$$

Apply (4.31) and (4.32) in the forms $\Lambda_{11}G_{11} = I + g_{12}G_{21}$ and $\Lambda_{11}G_{12} = g_{12}G$ to the terms in the upper right corner yield the reduction to Λ_{21}. The terms in the lower right need only be rearranged and the form of the transpo recognized. Q.E.D.

Theorem 4.1 can be used to structure the computation of M^{-1}. This computation is particularly easy when $\Lambda_{12} = 0 = \Lambda_{21}$. Then the main effort goes into producing G_{22} from (4.28). Otherwise alternative forms for the lower right submatrix of M^{-1} may be useful. Once G_{22} is made available, then the products $\Lambda_{21}G_{12}g_{22} = \Lambda_{21}\Lambda_{11}^{-1}g_{12}G_{22}$ using (4.26). Then G_{11} is obtained from the right number of (4.25). Assembling the results produces

$$g_{22}G_{22}g_{22} - \Lambda_{21}\Lambda_{11}^{-1}g_{12}G_{22}g_{22} - g_{22}G_{22}g_{21}\Lambda_{11}^{-1}\Lambda_{12} + \Lambda_{21}\Lambda_{11}^{-1}g_{12}G_{22}g_{21}\Lambda_{11}^{-1}\Lambda_{12} + \Lambda_{21}\Lambda_{11}^{-1}\Lambda_{12} \tag{4.23}$$

for this pesky submatrix.

Corollary 4.1. If $p = 2$ and $q = 1$ then

$$M^{-1} = \begin{bmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{12} & \dfrac{\Lambda_{11}g_{22}^2 - 2g_{12}g_{22}\Lambda_{12} + C_{22}\Lambda_{12}^2}{\Lambda_{11}G_{22} - g_{12}^2} \end{bmatrix}$$

proof. Under this hypothesis all the submatrices of M^{-1}, C, A are scalars. Using (4.22) and (4.28) one obtains $G_{22} = \Lambda_{11}/|C|$ where $|C| = \det C$, which when used in (4.33) will verify the lower right corner after reduction.

The matrices Λ and M^{-1} have the substructure of (4.1). By Lemma 4.1 the rank of $\Lambda - M^{-1}$ is no more that p-q and will be equal to p-q if, according to Lemma 4.2, the following matrix is invertible:

$$\begin{aligned} &\Lambda_{22} - \Lambda_{21}G_{11}\Lambda_{12} + \Lambda_{21}G_{12}g_{22} + (\Lambda_{21}G_{12}g_{22})' \\ &\quad - g_{22}G_{22}g_{22} \end{aligned} \tag{4.34}$$

Use of (4.33) can be made to express (4.34) as

$$\begin{aligned} &\Lambda_{22} - \Lambda_{21}\Lambda_{11}^{-1}\Lambda_{12} \\ &- [g_{22} - g_{21}\Lambda_{11}^{-1}\Lambda_{12}]' \, G_{22}[g_{22} - g_{21}\Lambda_{11}^{-1}\Lambda_{12}] \end{aligned} \tag{4.35}$$

One might expect that the characterization presented by Theorem 4.1 can be generalized to two different estimation systems, $g_1 = 0$ and $g_2 = 0$, that have some equations, $\mu = 0$, in common. That is, neither system is assumed to be the maximum likelihood system. In such a generalization, one would expect the two inverse covariance matrices, M_1^{-1} and M_2^{-1}, to have submatrices in common. In general this is false. A counter example, involving the gamma distribution, is presented in [15, Sec. 4]. The fundamental reason is that Lemma 4.4 is not available.

V. Directional Efficiency

Combining (3.7) and (4.14) lead to the multiparameter definition of asymptotic efficiency

$$\mathrm{Eff}(\tilde{\theta}) = \frac{1}{|M\Lambda|} \tag{5.1}$$

Since maximum likelihood estimates are efficient, consistent, and asymptotically normal, that is

$$\sqrt{n}\,(\hat{\theta}-\theta) \xrightarrow{\mathcal{L}} N(0,\Lambda^{-1}) \tag{5.2}$$

the expression (5.1) represents the comparative rate at which squared volumes of the ellipsoids of concentration of $\tilde{\theta}$ converges to zero. That is, the matrix of that ellipse is roughly $\frac{1}{n}M^{-1}$, and $|\Lambda^{-1}|$ represents the best rate at which it can shrink to zero as $n \to \infty$.

The estimate $\tilde{\theta}$ has been presented as a surrogate for $\hat{\theta}$, the maximum likelihood estimate, which may be too hard to find. The efficiency of $\tilde{\theta}$ will depend on two choices: the number q of likelihood equations retained and the quality of the replacement equations. The following concept of directional efficiency may be useful in examining these choices. The quantities $v'\hat{\theta}$ and $v'\tilde{\theta}$ are competing estimators of the linear combination $v'\theta$. The vector v specifies a direction in the parameter space. The

$$\lim_{n \to \infty} \frac{\mathrm{Var}(v'\hat{\theta})}{\mathrm{Var}(v'\tilde{\theta})} = \frac{v'\Lambda^{-1}v}{v'Mv} = e \tag{5.3}$$

is the (one-dimensional) asymptotic efficiency in the direction v.

The invariance feature of maximum likelihood estimates insures that $0 \le e \le 1$ for all v. Let us relate these directional efficiencies to the multivariate efficiency.

The right portion of (5.3) may be rewritten as

$$v'(eM - \Lambda^{-1})v = 0 \tag{5.4}$$

and serves to define e implicitly as a function of v. Since $\tilde{\theta}$ is asymptotically unbiased, we know from (2.5) and (2.28) that $N = I$ and

$$v'(M - \Lambda^{-1})v \geq 0 \tag{5.5}$$

for all directions v. Thus the directional efficiency e tells us how much M must be scaled down in order to produce zero in the direction v.

To characterize the critical values of $e(v)$ let us set the gradient of e equal to zero. This results in the system

$$(e(v)M - \Lambda^{-1})v = 0 \tag{5.6}$$

Since $v \neq 0$ and $e(v)$ is a scalar it is seen that there are p solutions (with possible multiplicities) to the equation $|eM - \Lambda^{-1}| = 0$. These critical solutions obviously satisfy

$$1 \geq e_1 \geq e_2 \geq \cdots \geq e_p > 0 \tag{5.7}$$

and, by the theory of simultaneous reduction of two quadratic forms [12], to each e_i is associated a critical direction v_i by solving (5.6) when $e(v)$ is set equal to e_i. It follows that

$$e_p \leq e(v) \leq e_1 \tag{5.8}$$

as v varies over the p-dimensional sphere and there exist directions of greater and lesser efficiency. Moreover, the critical efficiencies are the eigenvalues of $M^{-1}\Lambda^{-1}$, because (5.6) is equivalent to $(Ie - M^{-1}\Lambda^{-1})v = 0$. Upon multiplying them, it follows that

$$\prod_{i=1}^{p} e_i = \left|\frac{\Lambda^{-1}}{M}\right| = \mathrm{Eff}(\tilde{\theta}) \tag{5.9}$$

from (5.1). Thus the multiparameter efficiency is the product of the critical directional efficiences, and is at least as small as any of them.

Let us apply this material to that of Section IV. Suppose the first q equations of $g = 0$ are the first q likelihood equations. Clearly $|eM - \Lambda^{-1}| = 0$ is equivalent to

$$|e\Lambda - M^{-1}| = 0 \tag{5.10}$$

and the application of Theorem 4.1 shows that the root $e = 1$ appears at least q times. Thus, the directional efficiencies of $\tilde{\theta}$ are unity in a q dimensional subspace of Θ.

The above provides some quantification for the notion that q should be as close to p as possible. Turning to the question of measuring the quality of the replacement equations, the values of $e_{q+1},\ldots,e_p$ and their associated directions may prove useful.

We close with an example that illustrates the above features. Consider a gamma density

$$f(x;\alpha,\beta) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1}e^{-x/\beta}, \qquad 0 < x < \infty,\ 0 < \alpha,\ 0 < \beta \tag{5.11}$$

The properties (5.12) thru (5.16) are developed in [15, Sec. 4]: The maximum likelihood equations are

$$\bar{x} - \alpha\beta = 0 \tag{5.12}$$

$$\overline{\ln x} - \ln\beta - \psi(\alpha) = 0$$

where $\psi(\alpha)$ is the psi function. The information matrix is

$$\Lambda = \frac{1}{\beta^2}\begin{pmatrix} \alpha & \beta \\ \beta & \beta^2\psi'(\alpha) \end{pmatrix} \tag{5.13}$$

and the identification is $\theta_1 = \beta$, $\theta_2 = \alpha$. Since (5.12) cannot be solved explicitly, one often retreats to the method of moments, whose equations are (using s^2 for the sample variance)

$$\begin{aligned} \bar{x} - \alpha\beta &= 0 \\ s^2 - \alpha\beta^2 &= 0 \end{aligned} \tag{5.14}$$

which can be explicitly solved. Note that (5.14) shares an equation with (5.12) and Theorem 4.1 applies. In fact,

$$M^{-1} = \frac{1}{\beta^2} \begin{Bmatrix} \alpha & \beta \\ \beta & \dfrac{\beta^2}{\alpha} \dfrac{2\alpha + 3}{2(\alpha+1)} \end{Bmatrix} \tag{5.15}$$

Also, the efficiency of the method of moments is

$$\mathrm{Eff}(\tilde{\theta}) = \frac{1}{2(\alpha+1)\ (\alpha\psi'(\alpha)-1)} \tag{5.16}$$

Since $p = 2$ and $q = 1$ in this case, the root $e = 1$ appears once and the other root can be obtained using (5.16) in (5.9). Thus

$$e_1 = 1 \ , \qquad e_2 = \frac{1}{2(\alpha+1)(\alpha\psi'(\alpha)-1)} \tag{5.17}$$

and one can proceed to calculate the corresponding directions. The result is

$$v_1 = \binom{\alpha}{\beta} \ , \qquad v_2 = \binom{0}{1} \tag{5.18}$$

Full efficiency is available in the first direction, i.e. when estimating

$$v_1'\theta = 2\alpha\beta$$

or any scalar multiple thereof. Thus the product $\alpha\beta$ is efficiently estimated and this is not surprising. It is the quantity shared by the two systems (5.12) and (5.14).

The minimum efficiency is available in the second direction,

$$v_2'\theta = \alpha$$

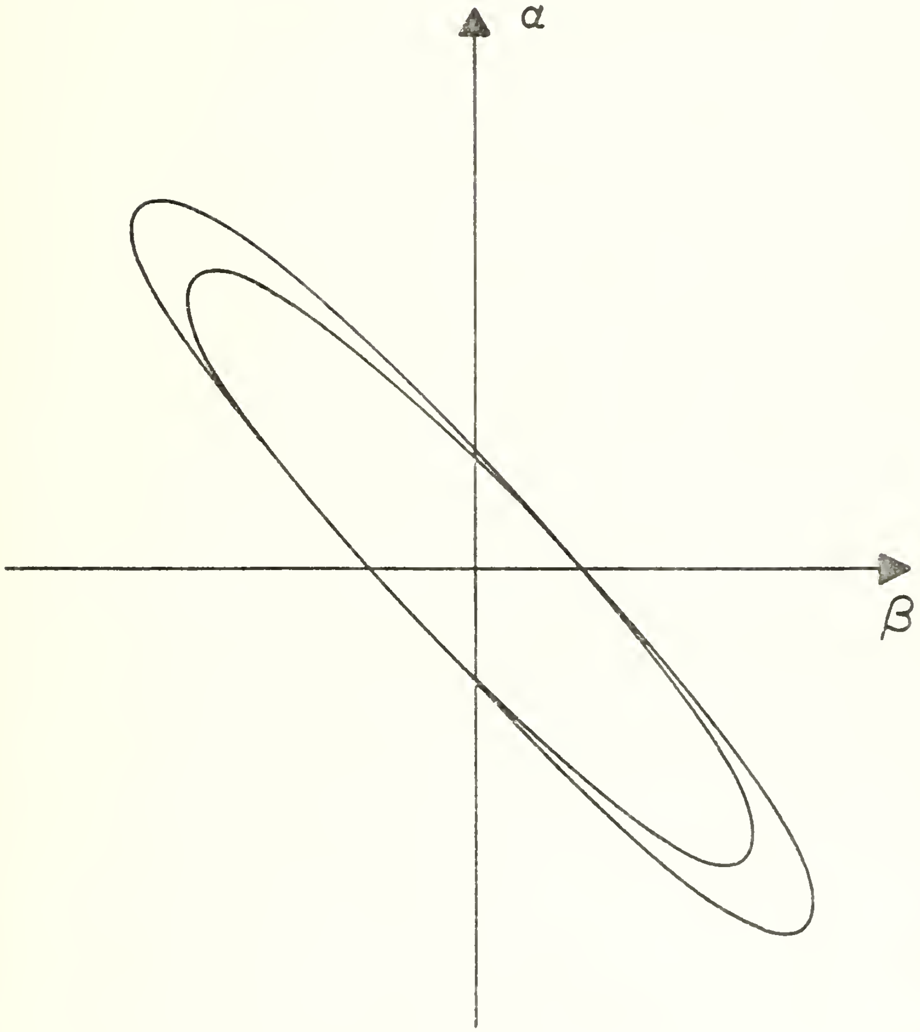
Thus the estimate of the shape parameter suffers most when (5.14) is used.

It seems useful to consider Figure 5.1, which contains the ellipses of concentration related to the maximum likelihood estimate (5.12) and the moment estimate (5.14). The inner ellipse corresponds to maximum likelihood and lies entirely within the outer since $M^{-1} - \Lambda \geq 0$. They come together and touch at the two points on the β axis, by Theorem 4.1. The marginal distributions obtained by projecting all of the probability mass onto a line of given direction will have maximum discrepancy if the projection takes place on the vertical axis. The two projections will coincide if done in the direction (α,β), e.g. 3, 2.5 in this case.

Figure 5.1

Ellipses of Concentration

(gamma: $\alpha = 3$, $\beta = 2.5$)



APPENDIX

Some Properties of the Geometric Distribution

The geometric distribution has density

$$f(x;p) = pq^{x} \qquad \text{for} \quad x = 0,1,\ldots \tag{A.1}$$

for a random variable X representing the number of failures preceding the first success in a series of Bernoulli trials. It is well known that

$$\mu = p/q \qquad\qquad \sigma^{2} = q/p^{2}$$

The maximum likelihood estimator for p is

$$\hat{p} = \frac{1}{1+X} \tag{A.3}$$

and its first two moments may be characterized as follows:

$$E(\hat{p}) = \sum_{0}^{\infty} \frac{1}{1+x}\, pq^{x} = \frac{p}{q} \sum_{0}^{\infty} \frac{q^{x+1}}{x+1} = \frac{p}{q} \sum_{0}^{\infty} \int_{0}^{q} u^{x} du$$

$$= \frac{p}{q} \int_{0}^{q} \sum_{0}^{\infty} u^{x} du = \frac{p}{q} \int_{0}^{q} \frac{du}{1-u} = -\frac{p}{q} \ln(p) \tag{A.4}$$

$$E(\hat{p}^{2}) = \sum_{0}^{\infty} \frac{1}{(1+x)^{2}}\, pq^{x} = \frac{p}{q} \sum_{0}^{\infty} \frac{q^{x+1}}{(x+1)^{2}} = \frac{p}{q} \sum_{0}^{\infty} \int_{0}^{q} \{ \int_{0}^{1} u^{x} du \} v^{x} dv$$

$$= \frac{p}{q} \int_{0}^{q} \int_{0}^{1} \frac{du\, dv}{1-uv} = -\frac{p}{q} \int_{0}^{q} \ln(1-v)\, \frac{dv}{v}$$

$$= \frac{p}{q} \int_{0}^{q} \sum_{1}^{\infty} \frac{v^{j-1}}{j} = \frac{p}{q} \sum_{1}^{\infty} \frac{q^{j}}{j^{2}} \tag{A.5}$$

The bias function is

$$b(p) = -\frac{p \ln p}{q} - p \tag{A.6}$$

which is uniformly positive and has derivative

$$b'(p) = \frac{-\ln(p)}{q^2} - \frac{1+q}{q} \tag{A.7}$$

which decreases monotonically and changes sign at about $p = .3162$. The maximum bias is .216. The log likelihood function has derivative

$$S_1 = \frac{1}{p} - \frac{X}{q} \tag{A.8}$$

and variance

$$E(S_1^2) = \frac{1}{qp^2} \tag{A.9}$$

To develop the sequence of Bhattacharyya bounds, one must first characterize (2.26)

$$S_r = \frac{1}{f(x;p)} \frac{\partial^r f(x;p)}{\partial p^r} \tag{A.10}$$

Using the notation

$$x^{(r)} = x(x-1) \cdots (x-r+1) \tag{A.11}$$

for factorials, one can verify

$$\frac{\partial^r f}{\partial p^r} = (-1)^{r-1} \frac{rx^{(r-1)}}{p} f(x-r+1) + (-1)^r x^{(r)} f(x-r) \tag{A.12}$$

by induction. Dividing (A.12) by (A.1) yields

$$S_r = \frac{(-1)^r}{q^r}\{X^{(r)} - \frac{rq}{p} X^{(r-1)}\} \tag{A.13}$$

and no end corrections are necessary. The covariance matrix calculations requires joint factorial moments.

Consider the generating function

$$G(u) = E(u^X) = \frac{p}{1-qu} \tag{A.14}$$

Differentiating r times allows

$$E(X^{(r)}) = \left.\frac{r!\, q^r p}{(1-qu)^{r+1}}\right|_{u=1} = r!\, (\frac{q}{p})^r \tag{A.15}$$

In similar fashion the product moments $X^{(s)}X^{(r)}$ can be obtained from

$$E(X^{(s)}X^{(r)}) = \left.\frac{\partial^{r+s} G(uv)}{\partial v^s\, \partial u^r}\right|_{u=v=1} \tag{A.16}$$

using the product argument uv in (A.14). One can proceed as follows. Starting with

$$\frac{\partial^r G(uv)}{\partial u^r} = \frac{r!\, pq^r v^r}{(1-quv)^{r+1}} \tag{A.17}$$

one can continue with the Newton differentiation formula

$$(fg)^{(s)} = \sum_{j=0}^{s} \binom{s}{j} f^{(j)} g^{(s-j)} \tag{A.18}$$

where in (A.18) the affixes refer to derivatives. For $j = 1,\ldots,s \leq r$ the jth derivative of v^r is

$$\frac{r!}{(r-j)!}\, v^{r-j} \tag{A.19}$$

and the (s-j)th derivative of $(1-quv)^{-r-1}$ is

$$\frac{(r+s-j)!}{r!}\,(qu)^{s-j}\,(1-quv)^{-r-1-s+j} \tag{A.20}$$

To collect: To obtain (A.16) one needs s derivatives of (A.17) with respect to v. These are provided by inserting (A.19) and (A.20) into (A.18) and setting $u = v = 1$. Thus for $s \leq r$

$$E(X^{(s)}X^{(r)}) = r!\,pq^r \sum_{j=0}^{s} \binom{s}{j} \frac{r!}{(r-j)!} \frac{(r-s-j)!}{r!} \frac{q^{s-j}}{(1-q)^{r+1+s-j}}$$

$$= r!\,s!\,(\tfrac{q}{p})^{r+s} \sum_{j=0}^{s} (\tfrac{p}{q})^{j} \binom{s}{j} \binom{r+s-j}{s} \tag{A.21}$$

which appears to be the most convenient form.

Returning to the question of product moments for (A.13), express

$$S_s S_r = \frac{(-1)^{r+s}}{q^{r+s}} \{X^{(r)}X^{(s)} - \frac{q}{p}\,[rX^{(r-1)}X^{(s)} + sX^{(r)}X^{(s-1)}] + rs(\tfrac{q}{p})^2\, X^{(r-1)}X^{(s-1)}\} \tag{A.22}$$

Applying the expectation operator to this using (A.21) produces, for $s < r$

$$E(S_sS_r) = \frac{(-1)^{r+s}\ r!\ s!}{p^{r+s}} \{\sum_0^s \left(\frac{p}{q}\right)^j \binom{s}{j} \binom{r+s-j}{r-j} - \sum_0^s \left(\frac{p}{q}\right)^j \binom{s}{j}\binom{r-1+s-j}{r-1-j}$$

$$- \sum_0^{s-1} \left(\frac{p}{q}\right)^j \binom{s-1}{j}\binom{r+s-1-j}{r-j} + \sum_0^{s-1} \left(\frac{p}{q}\right)^j \binom{s-1}{j}\binom{r+s-2-j}{r-1-j}\}$$

$$\text{(A.23)}$$

This can be reduced by applying the formula $\binom{n}{x-1} + \binom{n}{x} = \binom{n+1}{x}$ appropriately. Thus

$$\binom{r+s-j}{r-j} - \binom{r-1+s-j}{r-1-j} = \binom{r-1+s-j}{r-j}$$

and

$$\binom{r+s-1-j}{r-j}\ \binom{r+s-2-j}{r-1-j} = \binom{r+s-2-j}{r-j}.$$

can be inserted into that part of (A.23) which is enclosed in braces, producing the intermediate form

$$\sum_{j=0}^{s-1} \left(\frac{p}{q}\right)^j \left[\binom{s}{j}\binom{r-1+s-j}{r-j} - \binom{s-1}{j}\binom{r+s-2-j}{r-j}\right] + \left(\frac{p}{q}\right)^s \left[\binom{r}{r-s} - \binom{r-1}{r-1-s}\right] \quad \text{(A.24)}$$

The differences of products of binomial coefficients can be combined in a straightforward way so that (A.24) becomes

$$\sum_0^{s-1} \left(\frac{p}{q}\right)^j \binom{s}{j} \binom{r+s-1-j}{s-1} \frac{rs-j}{s(r+s-1-j)} + \left(\frac{p}{q}\right)^2 \binom{r-1}{r-s}\ .$$

Then one assembles the final form, for $s \leq r$

$$E(S_sS_r) = \frac{(-1)^{r+s}}{p^{r+s}}\ r!s! \sum_{j=0}^{s} \left(\frac{p}{q}\right)^j \binom{s}{j}\binom{r+s-1-j}{s-1} \frac{rs-j}{s(r+s-1-j)} \qquad \text{(A.25)}$$

The product moments (A.25) are covariances since each $E(S_r) = 0$. Because of the structure of (2.35), the bound L_k is the leading element of Λ^{-1}. Let us use the notation $\lambda_{rs} = E(S_r S_s)$ and apply (A.25) to show

$$\lambda_{1r} = \frac{(-1)^{r+1}\, r!}{qp^{r+1}} \qquad \text{for } r \geq 1$$

$$\lambda_{2r} = \frac{(-1)^{r+2}\, 2r!}{q^2 p^{r+2}} (r-1+q) \qquad \text{for } r \geq 2 \tag{A.26}$$

and that, for $k = 3$,

$$\Lambda = \begin{Bmatrix} \frac{1}{p^2 q} & \frac{-2}{p^3 q} & \frac{6}{p^4 q} \\ \frac{-2}{p^3 q} & \frac{4(1+q)}{p^4 q^2} & \frac{-12(2+q)}{p^5 q^2} \\ \frac{6}{p^4 q} & \frac{-12(2+q)}{p^5 q^2} & \frac{36(1+4q+q^2)}{p^6 q^3} \end{Bmatrix} \tag{A.27}$$

Direct calculations show that the cofactor of λ_{11} is

$$\frac{144}{p^{10} q^5} (1+q+q^2)$$

and that

$$|\Lambda| = \frac{144}{p^{12} q^6}$$

implying the desired form for L_3.

REFERENCES

[1] Abramowitz, M. and Stegun, I.A., Handbook of Mathematical Functions, National Bureau of Standards, Applied Mathematics Series, 55 (1969).

[2] Bhattacharyya, A., "On Some Analoges of the Amount of Information and Their Use in Statistical Estimation, Sankya 8 (1946), p. 1.

[3] Chapman, D.G. and Robbins, H., "Minimum Variance Estimation Without Regularity Assumptions," Annals of Math. Statis. 22 (1951), pp. 581-58

[4] Cramér, H., "A Contribution to the Theory of Statistical Estimation," Skand. Aktuarietidskrift 29 (1946), p. 85

[5] Cramér, H., Mathematical Methods of Statistics, Princeton (1951), pp. 489-494.

[6] Daniels, H.E., "The Asymptotic Efficiency of a Maximum Likelihood Estimator," Fourth Berkeley Symposium on Mathematical Statistics and Probability, Vol. I, University of California Press (1961).

[7] Graybill, F.A., Introduction to Matrices with Applications in Statistics, Wadsworth (1969).

[8] Kendall, M.G., The Advanced Theory of Statistics, Vol. II, Hafner, 195

[9] Kullback, S., Information Theory and Statistics, Wiley, 1959.

[10] LeCam, L.M., "On the Asymptotic Theory of Estimation and Testing Hypotheses," Third Berkeley Symposium on Mathematical Statistics and Probability, Vol. I, University of California Press, 1956.

[11] Lehman, E.L., "Notes on the Theory of Estimation," University of California, Department of Statistics, 1950.

[12] Mirksy, L., Introduction to Linear Algebra, Oxford, 1955.

[13] Rao, C.R., "Information and the Accuracy Attainable in the Estimation of Statistical Parameters," Bull. Calcutta Math. Soc. 37, No. 3 (1945), p. 81.

[14] Rao, C.R., "Asymptotic Efficiency and Limiting Information," Fourth Berkeley Symposium on Mathematical Statistics and Probability, Vol. I, University of California Press, 1961.

[15] Read, R.R., "Asymptotic Efficiency and Some Quasi-Method of Moments Estimators," Naval Postgraduate School, Technical Report NPS55-77-7, Monterey, Ca., 1977.

[16] Sokolnikoff, I.S., Advanced Calculus, McGraw-Hill, 1939.

[17] Wilks, S.S., Mathematical Statistics, Wiley, 1962.

INITIAL DISTRIBUTION LIST

	No. Copies
Defense Documentation Center Cameron Station Alexandria, Virginia 22314	2
Dean of Research Code 023 Naval Postgraduate School Monterey, California 93940	1
Library (Code 0212) Naval Postgraduate School Monterey, California 93940	2
Library (Code 55) Naval Postgraduate School Monterey, California 93940	2
Professor L. M. LeCam Department of Statistics University of California Berkeley, Ca. 94720	1
Department of Operations Research Naval Postgraduate School Monterey, Ca. 93940	
Attn: Professor D. R. Barr	1
Professor G. H. Bradley	1
Professor G. G. Brown	1
Professor J. Esary	1
Professor D. Gaver	1
Professor H.J. Larson	1
Professor P. Lewis	1
Professor K. Marshall	1
Professor P. Milch	1
Professor R. Read	20
Professor M. Sovereign	1
Professor P. Zehna	1
Professor Dan Davis Mathematics Department Naval Postgraduate School Monterey, California 93940	1